Analysis of Speaking Styles by Two-Dimensional Visualization of Aggregate of Acoustic Models

Makoto Shozakai and Goshu Nagino

Information Technology Laboratory
Asahi Kasei Corporation
{makoto, g-nagino}@ljk.ag.asahi-kasei.co.jp

Abstract

To ensure high enough recognition performance from the outset of usage of the speech recognition system, prior development of highly precise acoustic model library is necessary. The analysis of HMM acoustic models expressed with Gaussian distributions of multidimensional vectors is typically a difficult task. The COSMOS (aCOustic Space Map Of Sound) method featuring the visualization of distributions of the acoustic models in a two dimensional space by utilizing multidimensional scaling technique is proposed in order to support the analysis through capability of human visual perception. The effectiveness of the proposed technique is reviewed based on an analysis on speaking styles. The marginal region within the two-dimensional visual map (called COSMOS map) obtained by the proposed method contains acoustic models with lower recognition performance. It is possible to improve recognition performance by dividing the marginal region into several smaller zones in which separate acoustic model is trained and provided to the speakers belonging to the same zone.

1. Introduction

Demand for ASR(Automatic Speech Recognition) is increasing in embedded appliances such as vehicle navigation systems, personal digital assistants for mobile workers without sophisticated means of data entry, humanoid robots, and environmental control equipments for the disabled and the elderly. Although practical application of HMM-based speaker-independent ASR software is accelerating, such software is hardly universal in providing high speech recognition performance for every user. The performance of ASR systems tend to vary depending on diverse factors such as physical characteristics of individual persons(vocal cords, vocal tracts and etc.), task difficulties(size of vocabulary, balance of phonemes and etc.), speaking styles, ambient noise(loudness, type and etc.). In actual usage, the complex combination of these variables renders it extremely difficult to forecast and assure highly uniform level of recognition performance for every user. Previous studies in ASR aimed to structure highly complex acoustic models applicable to all conceivable circumstances. However, efforts in this direction are now running up against performance limitations. No matter how extensive the collected speech corpus is, or how sophisticated the acoustic model is, a uniformly high voice recognition performance is not provided to all users yet.

An alternative strategy, that is, speaker-adaptation approach was searched to overcome the current performance limitations. The speaker-adaptation approaches as represented by the MAP method [1] and the MLLR method [2] call for a sufficient amount of voice samples in order to take effect. However, it is often difficult to secure a sufficient amount of voice samples in practical use. The user finds it extremely troublesome and unaffordable to consume a significant amount of his/her time to supply the system with an applicable quantity of voice samples. The typical user naturally assumes the ASR system will live up to his/her expectation from the moment the system is activated for the first time. The ASR system vendors are therefore obligated to provide each and every user with an acceptable recognition performance from the beginning. Under the current circumstances, speaker-adaptation technology is only effective for users motivated to invest their time in providing an adequate amount of voice samples if such investment is to produce acoustic model optimized to the individual.

An approach featuring multiple acoustic models corresponding to the variables provided above should be effective in practical applications constrained by the difficulties of collecting an adequate quantity of voice samples for adaptation, enabling the user to select an appropriate acoustic model depending on the circumstance. This is analogous to the fitting and purchasing of pre-fabricated clothing most suited to the physical size and personal preferences of the user. It would be desirable to devise the ASR systems to allow for the selection of acoustic model by the user corresponding to the circumstance of usage. Within this paper, Section 2 below describes the proposed COSMOS method of visualizing the distribution of acoustic models in two-dimensional space to facilitate the understanding expanse of the acoustic space. In Section 3, one example of analysis on acoustic space containing variety of speaking styles is discussed. Finally, the conclusion of this paper is presented in Section 4.

2. COSMOS method

From the perspectives of manufacturers of the embedded appliances with voice recognition functionality, the current ASR software with constraints in the prediction and assurance of performance should seem less inviting than conventional methods of input such as keyboards, mouse and switches. In order for the ASR system to attain a higher level of reliability for each and every user, capabilities to supply and maintain acoustic models optimized to respective users must be established.

Methods currently proposed for the configuration of acoustic models with voice samples of multiple users clustered into individual speaker groups include the speaker clustering method [3], cluster adaptive training method [4], EigenVoice(EV) method [5] and so forth. However, it has been difficult to perceptually analyze results of speaker categorization obtained by these methods because everything...
of the clustering process is executed in invisible multidimensional space on computers. In order to efficiently increase the precision of acoustic models, it is essential to comprehend the overall configuration of the acoustic space of the input(acoustic features) consisting of voice and noise to be processed by the ASR system. To this end, the visualization of multidimensional information by mapping it onto a space of lower order utilizing a method referred to as Multidimensional Scaling(MDS) [6] is extremely effective. Without exception, the techniques shown in [6] utilize two-dimensional projections of the multidimensional vector information, and thus are useless in the mapping of information consisting of multidimensional Gaussian distributions. The technique based on the Principal Component Analysis(PCA) [7], however, suggests a method of mapping aggregations of concatenated vectors configured by the mean vectors of all phonemes utilizing primary and secondary principal components. However, the cumulative proportion of the primary and secondary principal components is only around 9 percent and is significantly lower than 80 percent that is normally required as cumulative proportion for the PCA. The resulting scatter diagram can hardly be considered as an accurate reproduction of the spatial information of the original multidimensional Gaussian distributions. The PCA-based approaches are not suitable for the two-dimensional visualization of acoustic models which incorporate three or more competing principal components. A procedure for mapping information containing multidimensional Gaussian distributions onto two-dimensional space without causing losses in information needs to be devised.

Although the acoustic feature parameter vectors may be utilized directly as a possible source of multidimensional information, but this approach is not practical given the vast number of data that must be processed. Within the scope of the studies conducted, the acoustic model was regarded as an approximation of the acoustic space. A multidimensional scaling method is devised so that it allows for the effective analysis of the acoustic space by making use of human visual perception of the distribution of acoustic models visualized as a two-dimensional space.

The method is proposed as an extension of the Sammon method [8] enabling a nonlinear projection of HMM-based acoustic models onto a two-dimensional space. The Sammon method is a technique of nonlinear projection featuring the optimization of mapped position coordinates within lower order dimensions by the steepest descent method, thereby minimizing the difference between the summation of the mutual distances among the multidimensional information existing within the multidimensional space and the summation of the mutual Euclidean distances of the mapped position coordinates existing in the lower order space. Within this projection, a pair of higher order information existing in close proximity is mapped relatively close to each other in the lower order reproduction, while a pair of higher order information with a greater mutual distance is placed further away in the lower order projection as well.

In general, an acoustic model is an aggregate of various acoustic units such as words, syllables, phonemes, diphones or others. Accordingly, the mutual distance $D(i, j)$ between acoustic model $i$ and acoustic model $j$ is defined by the following:

$$D(i, j) \equiv \sum_{k=1}^{K} d(i, j, k) * w(k) / \sum_{k=1}^{K} w(k)$$

(1)

whereas $d(i, j, k)$ denotes the mutual distance between the acoustic unit $k$ within the acoustic model $i$ and the acoustic unit $k$ in the acoustic model $j$. $w(k)$ represents occurrence frequency for the acoustic unit $k$. $K$ indicates total number of acoustic units. The resulting two-dimensional representation itself is called the COSMOS map, while the proposed mapping method is referred to as the COSMOS method. Respective acoustic model projected onto the COSMOS map is called STAR.

The following is an example of the application of the COSMOS method to the ASR middleware VORERO [9]. Designed and developed for embedded applications, the VORERO middleware incorporates diphone-level acoustic models of HMM based on single Gaussian distribution in order to reduce the required processing power and memory consumption.

Although publicly acknowledged distance measures including Euclidean distance between mean vectors of the Gaussian distributions and the Bhattacharyya distance are also available, the Euclidian distance of mean vectors normalized by variance vectors shall be adopted as $d(i, j, k)$ within this paper. Assuming all acoustic models share a common topology with a one-on-one state alignment between respective acoustic models, $d(i, j, k)$ may be expressed using the following equation.

$$d(i, j, k) = \frac{1}{S(k)} \sum_{s=0}^{n(k)-1} \frac{1}{L} \sum_{l=0}^{L} \left( \frac{\mu(i, k, s, l) - \mu(j, k, s, l)}{\sigma(i, k, s, l) + \sigma(j, k, s, l)} \right)^2$$

(2)

whereas $\mu(i, k, s, l)$ and $\sigma(i, k, s, l)^2$ denotes the mean and variance of dimension $l$ for the state $s$ of the acoustic unit $k$ within the acoustic model $i$. $S(k)$ represents the number of states of the acoustic unit $k$. $L$ signifies the number of dimension of the acoustic model. In the case of the VORERO middleware, the acoustic parameters consist of 10 MFCCs, 10 delta MFCCs, and 1 delta energy at sampling frequency of 11.025kHz. Therefore, $L$ equals 21.

3. Analysis of speaking styles

Variability based on the speaking styles of speech was investigated using the COSMOS method. 145 Japanese males uttered a list of 176 words taken from the phoneme balanced word set(called ATR5240 in Japan) in the three styles of speech indicated in Table 1.

All speech data (with the exception of the evaluation data described later) was used in the training of a speaking-style-independent/speaker-independent (SSI/SS) acoustic model. Then, this acoustic model was employed in turn as an initial model to retrain the speaking-style-dependent/speaker-dependent (SSD/SD) acoustic models using the Baum-Welch algorithm. Figure 1 shows the COSMOS for the SSD/SD STARs. The triangular STAR plotted in the center of Figure 1 signifies the SSI/SS acoustic model used as the initial model for retraining SSD/SD acoustic models. Two axes
respective indicating speed and volume are drawn by hand as bi-directional arrows in Figure 1.

Table 1: This is an example of a table

<table>
<thead>
<tr>
<th>Speaking style name</th>
<th>Instructions given to Japanese male speakers at recording</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Read utterance list at normal speed.</td>
</tr>
<tr>
<td>Fast</td>
<td>Read utterance list at faster-than-normal speed of speech.</td>
</tr>
<tr>
<td>High</td>
<td>Read utterance list at higher-than-normal tone of speech.</td>
</tr>
<tr>
<td>Whisper</td>
<td>Read utterance list at a level not to be drowned by near-by persons.</td>
</tr>
<tr>
<td>Loud</td>
<td>Read utterance list at a level to be heard by persons at some distance.</td>
</tr>
<tr>
<td>Lombard</td>
<td>Read utterance list among an ambient car noise.</td>
</tr>
<tr>
<td>Syllable-enhanced</td>
<td>Read utterance list by enhancing the Japanese syllable.</td>
</tr>
</tbody>
</table>

![Figure 1: Speaking-style COSMOS for Japanese males](image)

Then, the COSMOS was divided into five zones as shown in Figure 2. The segmentation was determined upon consideration of the distribution of the syllable-enhanced/speaker-dependent STARs in upper-left region of the COSMOS. Six zone acoustic models (Zone 1, Zone 2, Zone 3, Zone 4, Zone 5, Zone 2-5, Zone 1-5) were compiled using the five segmented zones and the STARs (with the exception of the evaluation data described later) distributed in all zones.

The STARs assigned with labels in Figure 2 were used for evaluation purposes. Table 2 indicates the recognition accuracy of the evaluation data in the six-zone acoustic models mentioned above. The evaluation data with an ambient exhibition-hall noise superimposed at a level of 20dB SNR was evaluated. The vocabulary size was 176 words. The highest accuracy for each evaluation data is in bold and underlined.

From the results indicated in Table 2, the following was assumed:

1. 75% of overall evaluation data achieved over 90% recognition accuracy for acoustic model of Zone 1-5. 80% of the evaluation data attained recognition accuracy in excess of 90% in cases where the acoustic model of the zone hosting the evaluation data or the optimal zone acoustic model was selected.

2. (2-1) The speaking-style ‘normal’ was dominant for Zone 1. All evaluation data residing in Zone 1 achieved over 97% in recognition accuracy when acoustic model of Zone 1 was employed.

2. (2-2) The speaking-style ‘whisper’ was dominant for Zone 2. All evaluation data located in Zone 2 achieved over 93% recognition accuracy when acoustic model of Zone 2 was employed.

![Figure 2: Divided zones corresponding to Figure 1](image)

Table 2: Recognition accuracies(%)
The speaking-style "syllable-enhanced" was dominant for Zone 3. The cluster of "syllable-enhanced" STARs is distinctively separated from the clusters of alternative speaking-styles. This is assumed to indicate the significant difference of the "syllable-enhanced" speaking-style from other speaking styles. Half of the evaluation data from Zone 3 fell short of the 90% recognition accuracy for acoustic model of Zone 1-5, but the recognition accuracy indicated a significant improvement when acoustic model of Zone 3 was used.

The speaking-styles "loud" and "i Lombardi sharing the similar acoustic features were dominant in Zone 4 as opposed to the distribution for the speaking-style "whisper" STARs. 75% of the evaluation data in Zone 4 provided for recognition accuracy of over 90% for acoustic model of Zone 4.

The speaking-styles "fast" and "high", which are closely located, were dominant in Zone 5. Residing at the opposite extreme of the STARs for the speaking-style "syllable-enhanced", 75% of the evaluation data belonging to the "fast" and "high" groups in Zone 5 achieved over 92% recognition accuracy for the acoustic model of Zone 5.

Figure 3 shows the estimated boundary curves for the achievement of over 90% recognition accuracy in Zones 2, 3, 4 and 5. The region enclosed in the dashed line and indicated with arrows represents the area assumed to provide for recognition accuracies of 90% or higher using respective zone acoustic model. It is expected for the ASR vender to widen the region enclosed in the dashed line.

Performance for acoustic model of Zone 2-5 was almost equivalent to that for acoustic model of Zone 1-5, suggesting that the speech data residing in Zone 1 has no influence over the precision of the acoustic model Zone 1-5. Speech data distributed along the external regions of the COSMOS seems to maintain a more profound significance than those of the inner region.

STARs situated in the inner region have higher recognition accuracy while the STARs distributed in the peripheral regions performed poorly. Thus, it is possible to predict the recognition performance of a STAR by determining its position within the COSMOS.

If the position of a STAR within the COSMOS can be determined by minimal speech sample of a new user, it will become possible to select the optimal acoustic model quickly, simply by selecting the acoustic model in the corresponding zone. Furthermore, this may enable more precise acoustic model to be designed by collecting speech data that was used to train the acoustic models of the STARs surrounding the position of the relevant STAR.

Understanding the overall acoustic space is facilitated significantly by generating groups of STARs and representing their spatial distribution visually within the two-dimensional space of the COSMOS. Observation of the clusters consisting of groups of STARs provides an effective approach to the design of acoustic model library. The proposed COSMOS method should offer a very powerful tool in evaluating not only the dependencies on speaking-style, but also in the evaluation of the susceptibility to task, noise, the effect of noise cancellation and distortion compensation techniques and other factors that affect speech recognition performance.

4. Conclusions

This paper proposes the COSMOS method for two-dimensional visualization of aggregated acoustic models. The analysis of the COSMOS for multiple speaking styles was implemented, resulting in a lower recognition performance for the groups of STARs situated in the peripheral regions of the map. Improvement of recognition performance is possible by further segmenting the peripheral regions into smaller zones, where individual zone acoustic models are trained. It was suggested that once the position of STAR is determined based on the minimal speech sample of a new user, it would be possible to design acoustic model by collecting speech data that was used to train the acoustic models of the STARs surrounding the position of the relevant STAR.

5. References